On the Finiteness of the Capacity of Continuous Channels

Jihad Fahs, Ibrahim Abou-Faycal
Dept. of Elec. and Comp. Engineering, American University of Beirut
Beirut 1107 2020, Lebanon
jjf03@aub.edu.lb, iaf@alum.mit.edu

***Abstract*— *Evaluating the channel capacity is one of many key problems in information theory. In this work we derive rather-mild sufficient conditions under which the capacity of continuous channels is* finite *and* achievable*.

These conditions are derived for generic, memoryless and possibly non-linear additive noise channels. The results are based on a novel sufficient condition that guarantees the convergence of differential entropies under point-wise convergence of probability density functions.

Perhaps surprisingly, the finiteness of channel capacity holds for the majority of setups, including those where inputs and outputs have possibly infinite second-moments.

I. Introduction

Over continuous-alphabets channels, a common belief is that with "sufficient" power, one is capable of transmitting at arbitrarily large rates. Stated differently, if an input of infinite power is allowed, the channel capacity is infinite. This belief is perhaps inspired from the well-known Additive White Gaussian Noise (AWGN) and linear Gaussian channels for example.

However, recent studies have suggested that for some channels this is not true: even if an infinite power input is allowed, the achievable rates are not arbitrarily large:

- In [1], the authors studied a linear additive-noise channel where the noise is heavy-tailed –modeled using alpha-stable statistics. They showed that even if the input constraint does allow for an infinite-power input, the channel capacity is finite. Actually, the authors found the optimal input to be surprisingly of finite power.
- In [2], the authors studied an additive Cauchy-distributed noise, and the constraints did allow as well for infinite-power input signals. The capacity was proven to be finite despite the fact that the optimal input was found in this case to have infinite power.

The natural question that arises is: "under which conditions does one have a finite channel capacity?", the answer to which does clearly depend on the input constraints, but also on the noise statistics. In this work, we study the interaction between the input constraints, the input-output relationship and the noise distribution and derive conditions on the triplet under which the channel capacity is finite.

This guarantee of finiteness is of high significance as it is typically the first step one would undertake in order to quantify the capacity of a channel at hand. Consider for example an additive Gaussian noise channel where the output Y is related to the input X as follows:

$$Y = X + N, \tag{1}$$

and where X is independent of the noise N. If no constraints are imposed on X, arbitrarily large transmission rates are achievable. If a second moment constraint is imposed instead, the capacity is finite. What if a "weaker" constraint is imposed on X. Could the rates be arbitrary large? For illustrative purposes, consider the "weaker" constraint $\mathsf{E}\left[\ln^2(1+|X|)\right] \leq A$ for some $A > 0$. This channel (1) is equivalent to the channel:

$$Y = \operatorname{sgn}(U)\left(e^{|U|} - 1\right) + N$$

where U now is average power constrained $\mathsf{E}\left[U^2\right] \leq A$. At a first look, it is not clear whether the capacity of such a channel is finite or not. Indeed, in some sense the channel is "exponentially amplifying" the input and by more than what the cost is constraining it. An appropriately-chosen Cauchy distributed input X will satisfy the constraint but will have an infinite second moment. The average of Y^2 will be infinite as well. Is the capacity of this channel finite? Our result provides an unexpected positive answer to this question.

Theoretical interests aside, it may seem unusual in a Gaussian setup to impose the constraint $\mathsf{E}\left[\ln^2(1+|X|)\right]$ or any other type of input constraints that permits $\mathsf{E}\left[X^2\right]$ to be infinite. However, when the channel model features noise distributions having an infinite second moment, as in the case of some channels subject to multiple access [3] or radio-frequency [4] interference, imposing a second moment constraint becomes less sensible; such a constraint masks the characterization of the behaviour of the transmission rates function of the quality of the channel since the channel signal-to-noise ratio will constantly evaluate to zero. Furthermore, we note that the usage of constraints allowing the input to have an infinite second moment has been previously examined within the context of robust estimation and detection theory [5]–[7].

More formally, the notion of capacity of a discrete memoryless channel was defined in the early works of Shannon [8], [9] to be "the largest" rate at which one can communicate over a channel with an arbitrarily low probability of error. Through a coding theorem, Shannon proved that the capacity is given by the solution to an optimization problem, whereby the mutual information between the input and output of the channel is maximized. When it comes to continuous channels the inputs of which are potentially constrained, the results were extended (see for example [9]–[11]) and the channel capacity was also tied to a constrained optimization problem.

This work was supported by AUB's University Research Board (URB) and the Lebanese National Council for Scientific Research (LNCSR).

Naturally, in both setups it is implicitly assumed that the optimization problem is "well-defined", for otherwise relating the channel capacity to a solution of a maximization of mutual information is problematic. In this work we tackle this assumption and provide a sufficient condition for such an optimization problem to be both *well-defined* and yielding a *finite* and *achievable solution* for a wide range of channels.

We consider a generic average-constrained channel model where the noise is additive and absolutely continuous. We prove in Section III that under very mild conditions on the noise and the constraint, the channel capacity is indeed finite and achievable.

We start by deriving sufficient conditions that ensure that mutual information is finite –and hence well-defined– and we make use of the extreme value principle [12] to ensure that the maximization problem yields a finite and achievable solution. This could be achieved by enforcing two characteristics:

1- The input space of feasible distribution functions is compact.
2- The mutual information between the input and the output of the channel is continuous in the input distribution function.

We emphasize that these two properties are intimately related to the channel model and the input constraints if any.

The generic model adopted in this work encompasses multiple channel models found in the literature: We consider input-output relationships that are possibly non-linear; A generic average cost function $\mathcal{C}(\cdot)$ is imposed on the input; The absolutely continuous additive noise has a finite "super-logarithmic moment"[†] as is the case for Gaussian, uniform, generalized Gaussian, generalized t, Pareto, Gamma, alpha-stable distributions, and their mixtures. We show that whenever the input cost function has a "super-logarithmic growth"[‡], the channel capacity is *finite* and *achievable*.

Establishing the continuity of mutual information under any "super-logarithmic" input constraint is achieved using a novel result on the *convergence of differential entropies*. While numerous studies have tackled this subject (see for example [13], [14]), the conditions presented in Section II are among the weakest that insure this convergence whenever Probability Density Functions (PDFs) converge point-wise.

The rest of the paper is organized as follows: In Section II, a preliminary theorem concerning the convergence of differential entropies is listed and proved. The primary problem and the main result are presented in Section III, where we describe the channel model and state the conditions under which our result holds. Technical proofs are derived in Section IV. Some extensions are listed in Section V and Section VI concludes the paper.

[†] A "super-logarithmic moment" is an expectation of the form $\mathsf{E}\left[f(|X|)\right]$ for some function $f(|x|) = \omega(\ln(|x|))$.
We say that $f(x) = \omega\left(g(x)\right)$ if and only if $\forall\,\kappa > 0, \exists\, c > 0$ such that $f(x) \geq \kappa g(x), \forall x \geq c$.

[‡] We say that a function $f(x)$ has a "super-logarithmic growth" whenever $f(|x|) = \omega(\ln(|x|))$.

II. Convergence of differential entropies

In this section we establish a sufficient condition for the convergence of differential entropies whenever there is point-wise convergence of the corresponding PDFs. More precisely, we prove a theorem that guarantees this convergence under some rather-mild sufficient conditions. In layman terms, this theorem states that whenever the PDFs satisfy a super-logarithmic type of moment, point-wise convergence will imply convergence of differential entropies. We emphasize that the new conditions are weaker than some of those derived by Godavarti et al. [14, Thm 1]. Alternative conditions found in [14, Thm 4] are not directly related to those presented hereafter.

Theorem 1. *Let the sequence of PDFs on $\mathbb{R}$, $\{p_m(y)\}_{m\geq 1}$ and $p(y)$ satisfy the following conditions:*

C1- *The PDFs $\{p_m(y)\}_m$ and $p(y)$ are uniformly upper-bounded:*

$$\exists\, M \in (0,\infty) \text{ s.t.} \sup_{y\in\mathbb{R}, m\geq 1} \Big\{ p_m(y), p(y) \Big\} \leq M. \quad (2)$$

C2- *There exists a non-negative and non-decreasing function $l : [0,\infty) \rightarrow [0,\infty)$, such that $l(y) = \omega\left(\ln(y)\right)$[†] and*

$$\sup_m \Big\{ \mathsf{E}_{p_m}\left[l(|Y|)\right], \mathsf{E}_p\left[l(|Y|)\right] \Big\} \leq L, \quad (3)$$

for some positive (finite) value L.

Under these conditions, $h(p_m) \rightarrow h(p)$ whenever the PDFs $p_m(y) \rightarrow p(y)$ point-wise.

Before we prove the theorem, we highlight the importance of condition C2 by providing an example where it is not satisfied, and the theorem does not hold.

Example 1. Consider the sequence of PDFs $\{p_m(x)\}_{m\geq 3}$ defined on $\mathbb{R}$ as follows:

$$p_m(x) = \begin{cases} 1 - \dfrac{1}{\ln m} & x \in [0;1] \\ \dfrac{1}{(\ln m)^2}\dfrac{1}{x} & x \in (1;m]. \end{cases}$$

This sequence of PDFs converges point-wise to $p(x)$, the uniform distribution on $[0,1]$, and condition C1 is satisfied with a uniform upperbound $M = 1$. Computing the differential entropies,

$$\begin{aligned} h(p) &= 0 \\ h(p_m) &= -\left(1 - \frac{1}{\ln m}\right)\ln\left(1 - \frac{1}{\ln m}\right) \\ &\qquad + \frac{2\ln(\ln m)}{(\ln m)^2}\int_1^m \frac{1}{x}\,dx + \frac{1}{(\ln m)^2}\int_1^m \frac{\ln x}{x}\,dx \\ &= -\left(1 - \frac{1}{\ln m}\right)\ln\left(1 - \frac{1}{\ln m}\right) + \frac{2\ln(\ln m)}{\ln m} + \frac{1}{2} \\ &\rightarrow \frac{1}{2} \text{ as } m \rightarrow \infty, \end{aligned}$$

and hence there is no convergence of differential entropies. This is explained by the fact that condition C2 is not satisfied. Indeed, consider any function $l(x)$ that is non-negative, non-decreasing and $l(x) = \omega\left(\ln x\right)$. By definition, for any $\kappa >$

0, there exists a $c > 0$ such that $l(x) \geq \kappa \ln x$ for $x \geq c$. Therefore, for any $m \geq c$,

$$\begin{aligned}
&\mathsf{E}_{p_m}\left[l(|X|)\right] \\
&= \left(1 - \frac{1}{\ln m}\right)\int_0^1 l(x)\,dx + \frac{1}{(\ln m)^2}\int_1^m \frac{1}{x}\,l(x)\,dx \\
&= \left(1 - \frac{1}{\ln m}\right)\int_0^1 l(x)\,dx + \frac{1}{(\ln m)^2}\int_1^c \frac{1}{x}\,l(x)\,dx \\
&\qquad\qquad + \frac{1}{(\ln m)^2}\int_c^m \frac{1}{x}\,l(x)\,dx \\
&\geq \left(1 - \frac{1}{\ln m}\right)\int_0^1 l(x)\,dx + \frac{1}{(\ln m)^2}\int_1^c \frac{1}{x}\,l(x)\,dx \\
&\qquad\qquad + \frac{\kappa}{(\ln m)^2}\int_c^m \frac{1}{x}\,\ln x\,dx \\
&= \left(1 - \frac{1}{\ln m}\right)\int_0^1 l(x)\,dx + \frac{1}{(\ln m)^2}\int_1^c \frac{1}{x}\,l(x)\,dx \\
&\qquad\qquad + \kappa\,\frac{(\ln m)^2 - (\ln c)^2}{2(\ln m)^2} \\
&\geq \kappa\,\frac{(\ln m)^2 - (\ln c)^2}{2(\ln m)^2},
\end{aligned}$$

which is greater than $\frac{3}{8}\kappa$ whenever $m > c^2$. Since the inequality holds for any $\kappa > 0$ and m large enough then $\sup_m \left\{\mathsf{E}_{p_m}\left[l(|X|)\right]\right\}$ is unbounded which violates condition C2. We proceed next to the proof of Theorem 1.

Proof: We start by noting that the differential entropies $h(p)$ and $\{h(p_m)\}_{m\geq 1}$ exist and are finite by virtue of the fact that the PDFs are upperbounded and have a finite logarithmic moment [15, Proposition 1].

Assume now that the conditions of the theorem hold and that p_m converges to p point-wise. If the upperbound (2) M is larger than one, consider the change of variables, $Z = MY$ (for which $h(Z) = h(Y) + \ln M$,) or equivalently the PDFs,

$$d(y) \hat{=} \frac{1}{M}p\left(\frac{y}{M}\right), \qquad d_m(y) \hat{=} \frac{1}{M}p_m\left(\frac{y}{M}\right), m \geq 1.$$

These densities are upperbounded by one and the sequence $\{d_m(y)\}$ converges point-wise to $d(y)$. Furthermore, the function $l'(y) = l(y/M)$ is non-negative, non-decreasing and $l'(y) = \omega\left(\ln(y)\right)$. Additionally,

$$\mathsf{E}_{d_m}\left[l'(|Y|)\right] = \mathsf{E}_{p_m}\left[l'(|MY|)\right] = \mathsf{E}_{p_m}\left[l(|Y|)\right] \leq L.$$

The conditions of the theorem therefore hold for the laws $\{d_m, d\}$ and in what follows we assume without loss of generality that $M \leq 1$, and the differential entropies are all non-negative .

Let $\tilde{y}$ be any positive scalar such that $l(\tilde{y}) > 0$, and denote by $q(y) = \frac{1}{\pi}\frac{1}{1+y^2}$ the Cauchy density. Then, using the convention "$0 \ln 0 = 0$" and the fact that $y \ln y \geq -\frac{1}{e}$ for $y > 0$, we can write

$$\begin{aligned}
&-\int\limits_{|y|\geq\tilde{y}} p(y)\ln p(y)\,dy \\
&= -\int\limits_{|y|\geq\tilde{y}} p(y)\ln q(y)\,dy + \int\limits_{|y|\geq\tilde{y}} q(y)\frac{p(y)}{q(y)}\ln\frac{q(y)}{p(y)}\,dy \\
&\leq \ln\pi\int\limits_{|y|\geq\tilde{y}} p(y)\,dy + \int\limits_{|y|\geq\tilde{y}} \ln\left[1+y^2\right]p(y)\,dy \\
&\qquad\qquad + \frac{1}{e}\int\limits_{|y|\geq\tilde{y}} q(y)\,dy \\
&\leq \frac{\ln\pi}{l(\tilde{y})}\int\limits_{|y|\geq\tilde{y}} l(|y|)\,p(y)\,dy + \int\limits_{|y|\geq\tilde{y}} \ln\left[1+y^2\right]p(y)\,dy \\
&\qquad + \frac{1}{e}\frac{1}{\ln\left[1+\tilde{y}^2\right]}\int\limits_{|y|\geq\tilde{y}} \ln\left[1+y^2\right]q(y)\,dy, \quad (4)
\end{aligned}$$

where equation (4) is due to the fact that $l(\cdot)$ is non-decreasing. Hence,

$$\begin{aligned}
&-\int\limits_{|y|\geq\tilde{y}} p(y)\ln p(y)\,dy \\
&\leq \ln\pi\frac{\mathsf{E}_p\left[l(|Y|)\right]}{l(\tilde{y})} + 2\int\limits_{|y|\geq\tilde{y}} \ln\left[1+|y|\right]p(y)\,dy \\
&\qquad + \frac{1}{e}\frac{\mathsf{E}_q\left[\ln\left[1+Y^2\right]\right]}{\ln\left[1+\tilde{y}^2\right]} \quad (5) \\
&\leq \frac{L\ln\pi}{l(\tilde{y})} + 2\sup_{|y|\geq\tilde{y}}\left\{\frac{\ln\left[1+|y|\right]}{l(|y|)}\right\}\int\limits_{|y|\geq\tilde{y}} l(y)\,p(y)\,dy \\
&\qquad + \frac{1}{e}\frac{\ln 4}{\ln\left[1+\tilde{y}^2\right]} \quad (6) \\
&\leq \frac{L\ln\pi}{l(\tilde{y})} + 2L\sup_{|y|\geq\tilde{y}}\left\{\frac{\ln\left[1+|y|\right]}{l(|y|)}\right\} + \frac{1}{e}\frac{\ln 4}{\ln\left[1+\tilde{y}^2\right]}, \quad (7)
\end{aligned}$$

where equation (5) is justified since $l(\tilde{y})$ is positive and $l(y)$ is non-negative. In order to write equation (6) we use the identity $\mathrm{E}_q\left[\ln\left(1+y^2\right)\right] = \ln 4$ [16, Sec.3.1.3, p.51]. The supremum in equations (6) and (7) is finite –and goes to 0– for $\tilde{y}$ large-enough because $l(y) = \omega\left(\ln y\right)$.

Since the upperbound (7) also holds for any $p_m(y)$, then for every $\delta > 0$, there exists a $\tilde{y} > 0$ such that for all $m \geq 1$:

$$\left|\int\limits_{|y|\geq\tilde{y}} p_m(y)\ln p_m(y)\,dy\right| < \delta \;\&\; \left|\int\limits_{|y|\geq\tilde{y}} p(y)\,\ln p(y)\,dy\right| < \delta.$$

It remains to show that

$$\lim_{m\to+\infty} -\int\limits_{|y|<\tilde{y}} p_m(y)\ln p_m(y)\,dy = -\int\limits_{|y|<\tilde{y}} p(y)\,\ln p(y)\,dy,$$

which is guaranteed by the Dominated Convergence Theorem (DCT) since $|p_m(y)\ln p_m(y)| \leq \frac{1}{e}$ by virtue of the fact that $p_m(y) \leq 1$ for all m, which completes the proof. ∎

III. Sufficient conditions for finiteness of channel capacity

In what follows we derive sufficient conditions for a memoryless additive-noise channel to have a finite and achievable capacity. More specifically, we consider a generic discrete-time real and memoryless noisy communication channel where the noise is additive and where the input and output are possibly non-linearly related as follows:

$$Y = f(X) + N, \tag{8}$$

where $Y \in \mathbb{R}$ is the channel output and where the input X is assumed to have an alphabet $\mathcal{X} \subseteq \mathbb{R}$. The channel's input is distorted according to the deterministic and possibly non-linear function $f(x)$. Additionally, the communication channel is subjected to an additive noise –that is independent of the input– that is absolutely continuous with PDF $p_N(\cdot)$.

Finally, we assume that the input is subject to an average cost constraint of the form: $\mathsf{E}\left[\mathcal{C}\left(|X|\right)\right] \leq A$, for some $A \in (0, \infty)$ and where $\mathcal{C}(\cdot)$ is some cost function:

$$\mathcal{C} : [0, \infty) \longrightarrow \mathbb{R}.$$

Accordingly, we define for $A > 0$

$$\mathcal{P}_A = \left\{ \text{Prob. distributions } F : \int \mathcal{C}\left(|x|\right)\, dF(x) \leq A \right\}, \tag{9}$$

the set of all distribution functions satisfying the average cost constraint.

The primary question that we would like to answer is whether one can reliably transmit an arbitrarily large number of bits per use over this channel. Said differently, are the achievable rates over this channel bounded? The answer to this question follows from those of the following two questions:

- Is the mutual information between a feasible input and the corresponding output always finite?
- If it is the case, can this mutual information be arbitrarily large?

A positive answer to the first question allows by the coding theorem [17] to state that the channel capacity is the supremum of the mutual information $I(\cdot)$ between the input X and output Y over all input probability distributions F that meet the constraint $\mathcal{P}_A$:

$$C = \sup_{F \in \mathcal{P}_A} I(F).$$

For the channel at hand (8), we note that the channel transition probability law is absolutely continuous with density function given by

$$p_{Y|X}(y|x) = p_N(y - f(x)), \quad y \in \mathbb{R},\ x \in \mathcal{X}. \tag{10}$$

and the mutual information may be expressed as [11]

$$I(F) \hat{=} \iint p_N\left(y - f(x)\right) \ln \left[\frac{p_N\left(y - f(x)\right)}{p(y;F)} \right] dy\, dF(x), \tag{11}$$

where $p(y;F) = \int p_N\left(y - f(x)\right)\, dF(x)$ is the output PDF.

Sufficient conditions

We make the following rather-mild assumptions:

- The cost function $\mathcal{C}(\cdot)$:
  - A1- The cost function is lower semi-continuous.
  - A2- The cost function is non-decreasing.
  - A3- $\mathcal{C}\left(|x|\right) = \omega\left(\ln |f(x)|\right)$.

  Without loss of generality, one may assume that $\mathcal{C}(\cdot)$ is non-negative. For if it were not, define $\mathcal{C}'(|x|) = \mathcal{C}(|x|) - \mathcal{C}(0)$ and adjust the input constraint accordingly.
- The function $f(\cdot)$:
  - A4- The function is continuous.
  - A5- The absolute value of the function $|f(\cdot)|$ is an non-decreasing function of $|x|$ and $|f(x)| \to +\infty$ as $|x| \to +\infty$.
- The noise PDF $p_N(\cdot)$:
  - A6- The PDF is continuous on $\mathbb{R}$.
  - A7- The PDF is upperbounded.
  - A8- There exits a non-decreasing function

    $$\mathcal{C}_N : [0, \infty) \longrightarrow \mathbb{R},$$

    such that $\mathcal{C}_N\left(|n|\right) = \omega\left(\ln |n|\right)$, and

    $$\mathsf{E}_N\left[\mathcal{C}_N\left(|N|\right)\right] = L_N < \infty.$$

    As an example, this condition holds true for any noise PDF whose tail is "faster" than $\frac{1}{x(\ln x)^3}$.

  Conditions A7 and A8 guarantee that the noise differential entropy h_N, exists and is finite [15, Proposition 1]. Since from an information theoretic perspective, the general channel model (8) is invariant with respect to output scaling, we consider without loss of generality that the noise PDF is upperbounded by one.

  Also without loss of generality, one may assume that $\mathcal{C}_N(\cdot)$ is non-negative. Otherwise, one may adopt $\mathcal{C}'_N(|x|) = \mathcal{C}_N(|x|) - \mathcal{C}_N(0)$.

The above assumptions are sufficient conditions on the triplet $f(\cdot)$, $\mathcal{C}(\cdot)$, and $p_N(\cdot)$ that guarantee the finiteness and the achievability of the capacity of channel (8):

Theorem 2. *Under conditions A1 through A8, the capacity of the average-cost constrained channel (8) is finite and achievable.*

Furthermore, the maximum is achieved by a unique F^* in $\mathcal{P}_A$ if and only if the output PDF is injective in F.

We point out that assumptions A4 through A8 are related to the channel model at hand and are not "conditions" per say. These assumptions are satisfied by the vast majority of common models found in the literature.

When thinking in terms of conditions on the input – controlled by the user, A1, A2 and A3 are to be considered. Note that these conditions are also common to all cost functions found in the literature. While A1 and A2 are rather technical, the relevance of A3 may be seen in the following example.

Example 2. Consider the linear additive channel (1), where now the noise N is a uniformly distributed random variable on the interval $[0, 1)$.

Let X_1 and X_2 be two discrete random variables taking integer values $k \geq 2$, with respective probability mass functions:

$$p_{X_1}(k) = B_1 \frac{1}{k(\ln k)^2}, \quad p_{X_2}(k) = B_2 \frac{1}{k(\ln k)^3}, \qquad k \geq 2,$$

where B_1 & B_2 are the normalizing finite constants,

$$B_1 = \left[\sum_{k=2}^{\infty} \frac{1}{k(\ln k)^2}\right]^{-1} \quad B_2 = \left[\sum_{k=2}^{\infty} \frac{1}{k(\ln k)^3}\right]^{-1}.$$

Let Y_1 and Y_2 be the outputs of channel (1) whenever its inputs are X_1 and X_2 respectively. Given the placement of the mass points, X_1 may be perfectly inferred from Y_1 and $H(X_1|Y_1) = 0$. Similarly $H(X_2|Y_2) = 0$ and therefore the mutual informations

$$I(X_1;Y_1) = H(X_1) - H(X_1|Y_1) = H(X_1)$$
$$I(X_2;Y_2) = H(X_2).$$

Computing $H(X_1)$ and $H(X_2)$, we obtain:

$$\begin{aligned} H(X_i) &= -\sum_{k\geq 2} p_{X_i}(k) \ln p_{X_i}(k) \\ &= -\ln B_i + B_i \sum_{k\geq 2} \frac{\ln k + (1+i)\ln(\ln k)}{k(\ln k)^{1+i}} \quad i = 1, 2, \end{aligned}$$

which diverges for $i = 1$ and converges for $i = 2$. Accordingly, the mutual information of channel (1) is infinite when the input is X_1 whereas it is finite for input X_2. Note that $\mathsf{E}[\ln X_1]$ is infinite while $\mathsf{E}[\ln X_2]$ is finite, and this example showcases the importance of condition A3 when it comes to the finiteness of mutual information. Whenever A3 is not enforced, the channel capacity might be infinite as X_1 yields an infinite mutual information. The theorem states that when the condition is enforced, the capacity will be finite.

An interesting observation is that both $\mathsf{E}\left[X_1^2\right]$ and $\mathsf{E}\left[X_2^2\right]$ are infinite, however as inputs to the channel they yield respectively an infinite and a finite mutual information. We proceed next to prove Theorem 2.

Proof: The first statement of the theorem is established using the extreme value principle which we state for completeness and can be found in [12]:

Theorem. *If $I(\cdot)$ is a* real-valued, *weak* continuous *functional on a* compact *set $\Omega \subseteq \mathcal{F}$, then $I(\cdot)$ achieves its maximum on Ω.*

In order to apply this principle, we show in Section IV that the set $\mathcal{P}_A$ is *compact* (Theorem 3) and that the mutual information $I(F)$ is *finite* and *continuous* (Theorems 4 and 5). Therefore, the capacity of the average-cost constrained channel is finite and achievable.

When it comes to uniqueness, since $\mathcal{P}_A$ is *convex* (Theorem 3) whenever $I(\cdot)$ is *strictly concave*, then the maximum

$$C = \max_{F \in \mathcal{P}_A} I(F),$$

is achieved by a unique F^* in $\mathcal{P}_A$.

Knowing that $I(\cdot)$ is *concave* (Theorem 5), its strict concavity is equivalent to the strict concavity of the output differential entropy in $p_Y(\cdot)$. This is indeed the case if and only if $p_Y(\cdot)$ is injective in F. ∎

The next section is dedicated to the proofs of Theorems 3, 4 and 5.

IV. Proofs of the Theorems

We use techniques that have been first developed in [11] and later adopted in various works on mutual information maximization as in [18]: Denote by $\mathcal{F}$ the space of all probability distribution functions on $\mathbb{R}$. We adopt weak convergence [19, III-1, Def.2, p.311] on $\mathcal{F}$, and use the Levy metric to metrize this weak convergence [5, Th.3.3, p.25]. The optimization is carried out in this metric topology.

Optimization set properties

Theorem 3. *Whenever conditions A1, A2, A3 and A5 are satisfied, the set $\mathcal{P}_A$ defined in (9) is convex and compact.*

Proof:

We note first that the theorem was shown to hold for cost functions of the form $\mathcal{C}(|x|) = |x|^r$, for $r > 1$ in [1], [18]. We adopt the same methodologies to generalize the results presented hereafter.

Convexity: Let F_1 and F_2 be two probability distribution functions in $\mathcal{P}_A$, and λ some scalar between 0 and 1. Define $F = \lambda F_1 + (1-\lambda)F_2$. It is clear that F is a probability distribution function because it is non-decreasing, right continuous, $F(-\infty) = 0$ and $F(+\infty) = 1$. Additionally,

$$\begin{aligned} \int_{\mathbb{R}} \mathcal{C}(|x|)\, dF &= \int_{\mathbb{R}} \mathcal{C}(|x|)\, d\left(\lambda F_1 + (1-\lambda)F_2\right) \\ &= \lambda \int_{\mathbb{R}} \mathcal{C}(|x|)\, dF_1 + (1-\lambda)\int_{\mathbb{R}} \mathcal{C}(|x|)\, dF_2 \\ &\leq \lambda A + (1-\lambda)A = A. \end{aligned}$$

Therefore, $F \in \mathcal{P}_A$ and $\mathcal{P}_A$ is convex.

Compactness: Consider a random variable X with probability distribution function $F \in \mathcal{P}_A$. Applying Markov's inequality to random variable $\mathcal{C}(|X|)$ yields,

$$\Pr\{\mathcal{C}(|X|) \geq \alpha\} \leq \frac{\mathsf{E}[\mathcal{C}(|X|)]}{\alpha}, \quad \forall \alpha > 0.$$

Now let

$$K = \inf\left\{x \in [0,\infty) \text{ s.t. } \mathcal{C}(x) \geq \alpha\right\} + 1,$$

which is always greater or equal to 1. For any finite value of α, such a K exists since $\mathcal{C}(x)$ increases to $+\infty$ as $x \to +\infty$ by virtue of properties A3 and A5. Additionally, since $\mathcal{C}(\cdot)$ is non-decreasing by property A2,

$$\begin{aligned} \Pr\{\mathcal{C}(|X|) \geq \alpha\} &\geq \Pr\left\{|X| > K-1\right\} \geq \Pr\{|X| \geq K\} \\ &\geq F(-K) + [1 - F(K)]. \end{aligned}$$

Hence, for all $F \in \mathcal{P}_A$, we obtain

$$F(-K) + [1 - F(K)] \leq \frac{\mathsf{E}[\mathcal{C}(|X|)]}{\alpha} \leq \frac{A}{\alpha}.$$

Therefore, for every $\epsilon > 0$, there exists a $K_\epsilon > 0$, namely

$$K_\epsilon = \min\left\{x \in [0, \infty) \text{ s.t. } \mathcal{C}(x) \geq (A/\epsilon)\right\} + 1,$$

such that

$$\sup_{F \in \mathcal{P}_A} \left[F(-K_\epsilon) + [1 - F(K_\epsilon)]\right] \leq \epsilon.$$

This implies that $\mathcal{P}_A$ is *tight* [19, III-2, Def.2, p.318]. By Phrokhorov's Theorem [19, III-2, Th.1, p.318], $\mathcal{P}_A$ is therefore relatively sequentially compact and every sequence $\{F_n\}$ of distribution functions in $\mathcal{P}_A$ has a convergent sub-sequence $\{F_{n_j}\}$ where the limit F^* does not necessarily belong to $\mathcal{P}_A$. If we prove that $F^* \in \mathcal{P}_A$, the latter will be sequentially compact and hence compact since the space is metrizable [20, Th.28.2, p.179]. In order to show that the limiting distribution function F^* is in $\mathcal{P}_A$, it must satisfy the cost constraint which is the case. In fact,

$$\int \mathcal{C}\left(|u|\right)\, dF^*(u) \leq \lim_{n_j \to \infty} \inf \int \mathcal{C}\left(|u|\right)\, dF_{n_j} \leq A,$$

where the first inequality holds because $\mathcal{C}(|u|)$ is lower semi-continuous (property A1), and is bounded from below by $\mathcal{C}\left(0\right)$ for all $u \in \mathbb{R}$ (property A2) [21, Th. 4.4.4]. In addition, the second inequality is valid since the sub-sequence $\{F_{n_j}\}$ is in $\mathcal{P}_A$ and therefore satisfies the cost constraint $\forall\, n_j$. Finally, $F^* \in \mathcal{P}_A$ and $\mathcal{P}_A$ is compact. ∎

Properties of the mutual information, $I(\cdot)$

We prove in what follows the finiteness, concavity and continuity of $I(\cdot)$ on $\mathcal{P}_A$ through Theorems 4 and 5.

Theorem 4. *Whenever conditions A3, A7 and A8 hold, the mutual information $I(F)$ between the input and output of channel (8) is finite for all input distribution functions F such that $\mathsf{E}\left[\mathcal{C}\left(|X|\right)\right]$ is* finite.

Proof: Since $Y = f(X) + N$,

$$\ln\left[1 + |Y|\right] \leq \ln\left[1 + |f(X)|\right] + \ln\left[1 + |N|\right],$$

and $\mathsf{E}\left[\ln\left[1 + |Y|\right]\right]$ is finite because both $\mathsf{E}\left[\ln\left[1 + |f(X)|\right]\right]$ and $\mathsf{E}\left[\ln\left[1 + |N|\right]\right]$ are finite (by properties A3 and A8).

Moreover, and since $p_Y(y)$ is upperbounded (by A7) by one $h_Y(F) = -\int p(y;F)\, \ln p(y;F)\, dy$, the differential entropy of Y, is well defined [15, Proposition 1] and $0 \leq h_Y(y) < +\infty$.

The differential entropy h_N of the noise being finite (due to properties A7 and A8), the mutual information (11) can therefore be written as the difference of two terms:

$$I(F) \,= h_Y(F) \,-\, h_{Y|X}(F) \,= h_Y(F) \,-h_N, \tag{12}$$

both of which are finite and this completes the proof. ∎

Theorem 5. *Assume that conditions A2 through A8 hold. Under a cost constraint*

$$\int \mathcal{C}\left(|X|\right)\, dF(x) \leq A \qquad A > 0,$$

the mutual information $I(F)$ between the input and the output of channel (8) is concave and continuous whenever $\mathcal{C}\left(|x|\right) = \omega\left(\ln|f(x)|\right)$.

Before we proceed with the proof, we note that under the conditions of the Theorem, the mutual information $I(F)$ between the input and the output of channel (8) is finite by virtue of Theorem 4.

Proof:

Concavity: The output differential entropy $h_Y(F)$ is a concave function of F on $\mathcal{F}$. In fact,

$$h_Y(F) \,=\, -\int p_Y(y;F)\, \ln\, p_Y(y;F)\, dy$$

exists (by Theorem 4) and is a concave function of $p_Y(\cdot)$ because $-x\ln x$ is concave in x. Since $p_Y(F)$ is linear in F, $I(F) \,= h_Y(F) \,-\, h_N$ is concave on $\mathcal{P}_A$.

Continuity: To prove the continuity of $I(F)$, it suffices to show that $h_Y(F)$ is continuous by virtue of equation (12). To this end, we let $F \in \mathcal{P}_A$ and let $\{F_m\}_{m\geq 1}$ be a sequence of probability measures in $\mathcal{P}_A$ that converges weakly to F.

In order to apply Theorem 1 and show the convergence of $h_Y(F_m)$ to $h_Y(F)$ and hence the weak continuity of $h_Y(F)$ on $\mathcal{P}_A$, we establish that the appropriate conditions are satisfied:

- By definition of weak convergence, since $p_N(y-x)$ is bounded and continuous (properties A6 and A7), then $p(y;F_m) \;=\; \int p_N(y \,-\, f(x))\, dF_m(x)$ converges point-wise to $p(y;F) = \int p_N(y - f(x))\, dF(x)$.
- The induced output PDF $p(y;F_m)$ is also bounded by one.
- It remains to find a non-negative and non-decreasing function, $l : [0,\infty) \to [0,\infty)$ such that $l(y) = \omega\left(\ln(y)\right)$, and a scalar $L > 0$ such that equation (3) holds for $p(y;F_m)$, $m \geq 1$ and $p(y;F)$, a task which we fulfill in what follows.

For any $y \geq |f(0)|$, let $\mathcal{S} = f^{-1}\left([|f(0)|, y]\right)$ be the inverse image by $f(\cdot)$ of the closed interval $[|f(0)|, y]$. Since $f(\cdot)$ is continuous (A4), the set $\mathcal{S}$ is closed. It is also bounded because $|f(x)|$ is non-decreasing in $|x|$ and tends to infinity (A5). Therefore any element in $\mathcal{S}$ is smaller than a positive t_u such that $|f(t_u)| = 2y$ and greater than a negative t_b such that $|f(t_b)| = 2y$. Such t_u and t_b exist because $f(\cdot)$ is continuous.

The set is compact and has a maximal value that we denote $z(y) = \max\{z : z \in \mathcal{S}\}$. Note that $|f(z(y))| = y$.

Define the function $\mathcal{C}_{\min}(\cdot)$: $[0, \infty) \longrightarrow \mathbb{R}$ as follows:

$$\mathcal{C}_{\min}\left(y\right) = \begin{cases} \min\left\{\mathcal{C}\left(z(y)\right); \mathcal{C}_N(y)\right\} & y \geq |f(0)| \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{C}_N(\cdot)$ is defined in A8. The function $\mathcal{C}_{\min}(y)$ is non-negative and non-decreasing on $[0, \infty)$ since both $\mathcal{C}(y)$ and $\mathcal{C}_N(\cdot)$ are non-negative and non-decreasing by properties A2 and A8 and $z(y)$ is non-decreasing for $y \geq |f(0)|$. Additionally, $\mathcal{C}_{\min}(y) \,=\, \omega\left(\ln y\right)$ because $\mathcal{C}(x) \,=\, \omega\left(\ln|f(x)|\right)$ (A3) and $\mathcal{C}_N(x) = \omega\left(\ln x\right)$ (A8).

Now, for any X with distribution $F \in \mathcal{P}_A$,

$$\begin{aligned}
&\mathsf{E}_Y\left[\mathcal{C}_{\min}\left[\frac{|Y|}{2}\right]\right] \\
&= \mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left[\frac{|f(X)+N|}{2}\right]\right] \\
&\leq \mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left[\frac{|f(X)|+|N|}{2}\right]\right] \qquad (13)\\
&= \mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left[\frac{|f(X)|+|N|}{2}\right]\Bigg|_{|f(X)|\leq|N|}\right]\mathrm{P}\left(|f(X)|\leq|N|\right) \\
&\quad+\mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left[\frac{|f(X)|+|N|}{2}\right]\Bigg|_{|f(X)|>|N|}\right]\mathrm{P}\left(|f(X)|>|N|\right) \\
&\leq \mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left(|N|\right)\Bigg|_{|f(X)|\leq|N|}\right]\mathrm{P}\left(|f(X)|\leq|N|\right) \\
&\quad+\mathsf{E}_{X,N}\left[\mathcal{C}_{\min}\left(|f(X)|\right)\Bigg|_{|f(X)|>|N|}\right]\mathrm{P}\left(|f(X)|>|N|\right) \qquad (14)\\
&\leq \mathsf{E}_N\left[\mathcal{C}_{\min}\left(|N|\right)\right]+\mathsf{E}_X\left[\mathcal{C}_{\min}\left(|f(X)|\right)\right] \qquad (15)\\
&\leq \mathsf{E}_N\left[\mathcal{C}_N\left(|N|\right)\right]+\mathsf{E}_X\left[\mathcal{C}\left(|X|\right)\right] \\
&\leq L_N + A = L < \infty. \qquad (16)
\end{aligned}$$

where $0 \leq L_N = \mathsf{E}_N\left[\mathcal{C}_N\left(|N|\right)\right] < \infty$ by property A8. Equations (13) and (14) are justified since $\mathcal{C}_{\min}\left(|x|\right)$ is non-decreasing in $|x|$ and (15) is due to the fact that $\mathcal{C}_{\min}\left(|x|\right)$ is non-negative. Since the value $0 \leq L < \infty$ is independent of the input distribution function $F \in \mathcal{P}_A$, then (16) holds for any output variable Y, i.e for all $p(y;F)$ where $F \in \mathcal{P}_A$. Letting $l(y) = \mathcal{C}_{\min}\left(\frac{y}{2}\right)$, $y \in [0,\infty)$, then equation (3) is satisfied for $p(y;F_m)$, $m \geq 1$ and $p(y;F)$. Therefore, Theorem 1 holds and $h_Y(F_m)$ converges to $h_Y(F)$ and hence $h_Y(F)$ is continuous which concludes the proof. ■

V. Extensions

The results may be extended to the case where the noise PDF is not necessarily continuous on $\mathbb{R}$. In fact, we weaken condition A6 and we show that Theorem 2 also holds for noise PDFs which are piece-wise continuous on a countable number of pieces. Note that under this category fall absolutely continuous noise variables with a compact support[††] such as the uniform, and also ones that are one-sided such as the Gamma or the Pareto random variables. We start by noting the following:

- It can be seen from the proof of Theorem 1 that almost everywhere (a.e.)[‡‡] point-wise convergence with respect to the Lebesgue measure (in addition to C1 and C2) is sufficient in order to have convergence of differential entropies.
- According to the definition of weak convergence [22, p. 700], one can replace continuous bounded test functions by F-a.e. continuous functions where F is the limit distribution.

We show now that if $p_N(\cdot)$ has a countable number of discontinuities then weak convergence of the input distributions implies Lebesgue-a.e. point-wise convergence of the output PDFs: Denote by $\{a_i\}_{i\geq 1}$ the countable discontinuities of $p_N(\cdot)$ and by $\{x_i\}_{i\geq 1}$ the discontinuity points of F, which are necessarily countable (see Jordan decomposition lemma in [23, p. 40]). Point-wise convergence of the PDFs holds except at values of y of the form $y_{ij} = a_i - f(x_j)$, $i,j \geq 1$. The fact that the $\{y_{ij}\}$'s are countable proves our assertion.

[††] we define the support of a random variable as being the set of its points of increase, i.e., the set $\{x \in \mathbb{R} : \Pr(x-\eta < X < x+\eta) > 0 \text{ for all } \eta > 0\}$.

[‡‡] We say that a property holds almost everywhere with respect to a measure μ and we denote it μ-a.e. if and only if the measure by μ of the set where the property fails is equal to zero.

VI. Conclusion

Tangible models for communication channels implicitly assume a finite value for the channel capacity. Knowing that maximizing the transmission rates is directly related to a constrained maximization problem, we have derived sufficient conditions for finiteness and achievability of the capacity of generic memoryless additive noise channels. The involved conditions on the input-output relationship, the input cost function and the type of the noise define a wide collection of models for which finding codes that strive toward achieving maximum transmission rates is sensible. The result is applicable to possibly non-linear channels, to nearly all the widely known additive noise models and for cost functions that are "super-logarithmic". Interestingly, communications at finite rates is not directly related to an input average-power constraint. Even when signaling strategies are allowed to have an infinite second moment on average, transmission rates could not be arbitrarily large. We mention that while searching for sufficiency, intermediately we derived conditions under which point-wise convergence of PDFs implies convergence of differential entropies.

References

[1] J. Fahs and I. Abou-Faycal, "On the capacity of additive white alpha-stable noise channels," in *IEEE International Symposium on Information Theory*, Cambridge, MA, USA, 2012, pp. 294–298.

[2] J. Fahs, I. Abou-Faycal, "A Cauchy input achieves the capacity of a Cauchy channel under a logarithmic constraint," in *IEEE International Symposium on Information Theory*, Honolulu, HI, USA, June 29 - July 4 2014.

[3] H. El Ghannudi, L. Clavier, N. Azzaoui, F. Septier, and P. a. Rolland, "Stable interference modeling and Cauchy receiver for an IR-UWB ad hoc network," *Communications, IEEE Transactions on*, vol. 58, no. 6, pp. 1748 –1757, Jun. 2010.

[4] M. Nassar, K. Gulati, A. Sujeeth, N. Aghasadeghi, B. Evans, and K. Tinsley, "Mitigating near-field interference in laptop embedded wireless transceivers," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 1405 –1408.

[5] P. J. Huber, *Robust Statistics*. John Wiley & Sons, 1981.

[6] M. Shao and C. Nikias, "Signal processing with fractional lower order moments: stable processes and their applications," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 986 –1010, Jul. 1993.

[7] J. G. Gonzalez, D. W. Griffith, and G. R. Arce, "Zero-order statistics: A signal processing framework for very impulsive processes," *Proceedings of the IEEE Signal Processing Workshop*, pp. 254–258, 1997.

[8] C. E. Shannon, "A mathematical theory of communication, part i," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[9] ——, "A mathematical theory of communication, part ii," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[10] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Nov. 1968.

[11] J. G. Smith, "The information capacity of peak and average power constrained scalar Gaussian channels," *Inform. Contr.*, vol. 18, pp. 203–219, 1971.

[12] D. G. Luenberger, *Optimization By Vector Space Methods*. New York: Wiley, 1969.

[13] F. J. Piera and P. Parada, “On convergence properties of Shannon entropy,” *Problems of Information Transmission*, vol. 45, no. 2, pp. 75–94, 2009.
[14] M. Godavarti and A. Hero, “Convergence of differential entropies,” *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 1–6, January 2004.
[15] O. Rioul, “Information theoretic proofs of entropy power inequality,” *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 33–55, January 2011.
[16] J. N. Kapur, *Maximum Entropy Models In Science And Engineering*. Wiley, 1989.
[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
[18] J. Fahs, I. Abou-Faycal, “Using Hermite bases in studying capacity-achieving distributions over AWGN channels,” *Information Theory, IEEE Transactions on*, vol. 58, no. 8, August 2012.
[19] A. N. Shiryaev, *Probability*, 2nd ed. Springer-Verlag, 1996.
[20] J. R. Munkres, *Topology*, 2nd ed. Prentice Hall, 2000.
[21] K. L. Chung, *A Course in Probability Theory*, 3rd ed. San Diego: Academic Press, 2001.
[22] A. M. Zapala, “Unbounded mappings and weak convergence of measures,” *Statistics and Probability Letters*, vol. 78, pp. 698–706, 2008.
[23] A. K. Basu, *Measure Theory and Probability*, 2nd ed. PHI Learning, 2012.